Method and Apparatus for Determining Latency Between Multiple Servers and a Client

ABSTRACT

A method and apparatus for determining latency between multiple servers and a client receives requests for content server addresses from local domain names servers (LDNS). POPs that can serve the content are determined and sent latency metric requests. The content server receives the request for latency metrics and looks up the latency metric for the requesting client. Periodic latency probes are sent to the IP addresses in a Latency Management Table. The IP addresses of clients are masked so the latency probes are sent to higher level servers to reduce traffic across the network. The hop count and latency data in the packets sent in response to the latency probes are stored in the Latency Management Table and is used to determine the latency metric from the resident POP to the requesting client before sending the latency metric to the requesting server. The BGP hop count in the Latency Management Table is used for the latency metric upon the first request for an IP address. The latency metric is calculated for subsequent requests of IP addresses using the hop count and RTT data in the Latency Management Table. Latency metrics from POPs are collected and the inverse relationship of the hop counts in a weighted combination with the RTT are used to determine which latency metric indicates the optimal POP. The address of the optimal POP is then sent to the requesting LDNS.